

## Master Thesis Defense

Speaker:	Malik Waqas Sagheer
Supervisor:	Dr. C. Y. Suen
Examining Committee:	Drs. T. D. Bui, D. Ford, N. Shiri (Chair)
Title:	Novel Word Recognition and Word Spotting Systems for Offline Urdu Handwriting
Date:	Wednesday, February 17, 2010
Time:	10:00 a.m.
Place:	EV 3.101

### ABSTRACT

Word recognition for offline Arabic, Farsi and Urdu handwriting is a subject which has attained much attention in the OCR field. This thesis presents the implementations of offline Urdu Handwritten Word Recognition (HWR) and an Urdu word spotting technique. This thesis first introduces the creation of several offline CENPARMI Urdu databases. These databases were necessary for offline Urdu HWR experiments. The holistic-based recognition approach was followed for the Urdu HWR system. In this system, the basic pre-processing of images was performed. In the feature extraction phase, the gradient and structural features were extracted from greyscale and binary word images, respectively. This recognition system extracted 592 feature sets and these features helped in improving the recognition results. The system was trained and tested on 57 words. Overall, we achieved a 97 % accuracy rate for handwritten word recognition by using the SVM classifier.

Our word spotting technique used the holistic HWR system for recognition purposes. This word spotting system consisted of two processes: the segmentation of handwritten connected components and diacritics from Urdu text lines and the word spotting algorithm. A small database of handwritten text pages was created for testing the word spotting system. This database consisted of texts from ten Urdu native speakers. The rule-based segmentation system was applied for segmentation (or extracting) for handwritten Urdu subwords or connected components from text lines. We achieved a 92% correct segmentation rate for 372 text lines.

In the word spotting algorithm, the candidate words were generated from the segmented connected components. These candidate words were sent to the holistic HWR system, which extracted the features and tried to recognize each image as one of the 57 words. After classification, each image was sent to the verification/rejection phase, which helped in rejecting the maximum number of unseen (raw data) images. Overall, we achieved a 50% word spotting precision at a 70% recall rate.